

# The genome of the cucumber, *Cucumis sativus* L.

Sanwen Huang<sup>1,19</sup>, Ruiqiang Li<sup>2,3,19</sup>, Zhonghua Zhang<sup>1,19</sup>, Li Li<sup>2,19</sup>, Xingfang Gu<sup>1,19</sup>, Wei Fan<sup>2,19</sup>, William J Lucas<sup>4,19</sup>, Xiaowu Wang<sup>1</sup>, Bingyan Xie<sup>1</sup>, Peixiang Ni<sup>2</sup>, Yuanyuan Ren<sup>2</sup>, Hongmei Zhu<sup>2</sup>, Jun Li<sup>2</sup>, Kui Lin<sup>5</sup>, Weiwei Jin<sup>6</sup>, Zhangjun Fei<sup>7</sup>, Guangcun Li<sup>8</sup>, Jack Staub<sup>9</sup>, Andrzej Kilian<sup>10</sup>, Edwin A G van der Vossen<sup>11</sup>, Yang Wu<sup>5</sup>, Jie Guo<sup>5</sup>, Jun He<sup>1</sup>, Zhiqi Jia<sup>1</sup>, Yi Ren<sup>1</sup>, Geng Tian<sup>2</sup>, Yao Lu<sup>2</sup>, Jue Ruan<sup>2,12</sup>, Wubin Qian<sup>2</sup>, Mingwei Wang<sup>2</sup>, Quanfei Huang<sup>2</sup>, Bo Li<sup>2</sup>, Zhaoling Xuan<sup>2</sup>, Jianjun Cao<sup>2</sup>, Asan<sup>2</sup>, Zhigang Wu<sup>2</sup>, Juanbin Zhang<sup>2</sup>, Qingle Cai<sup>2</sup>, Yinqi Bai<sup>2</sup>, Bowen Zhao<sup>13</sup>, Yonghua Han<sup>6</sup>, Ying Li<sup>1</sup>, Xuefeng Li<sup>1</sup>, Shenhao Wang<sup>1</sup>, Qiuxiang Shi<sup>1</sup>, Shiqiang Liu<sup>1</sup>, Won Kyong Cho<sup>14</sup>, Jae-Yean Kim<sup>14</sup>, Yong Xu<sup>15</sup>, Katarzyna Heller-Uszynska<sup>10</sup>, Han Miao<sup>1</sup>, Zhouchao Cheng<sup>1</sup>, Shengping Zhang<sup>1</sup>, Jian Wu<sup>1</sup>, Yuhong Yang<sup>1</sup>, Houxiang Kang<sup>1</sup>, Man Li<sup>1</sup>, Huiqing Liang<sup>2</sup>, Xiaoli Ren<sup>2</sup>, Zhongbin Shi<sup>2</sup>, Ming Wen<sup>2</sup>, Min Jian<sup>2</sup>, Hailong Yang<sup>2</sup>, Guojie Zhang<sup>2,12</sup>, Zhentao Yang<sup>2</sup>, Rui Chen<sup>2</sup>, Shifang Liu<sup>2</sup>, Jianwen Li<sup>2</sup>, Lijia Ma<sup>2,12</sup>, Hui Liu<sup>2</sup>, Yan Zhou<sup>2</sup>, Jing Zhao<sup>2</sup>, Xiaodong Fang<sup>2</sup>, Guoqing Li<sup>2</sup>, Lin Fang<sup>2</sup>, Yingrui Li<sup>2,12</sup>, Dongyuan Liu<sup>2</sup>, Hongkun Zheng<sup>2,3</sup>, Yong Zhang<sup>2</sup>, Nan Qin<sup>2</sup>, Zhuo Li<sup>2</sup>, Guohua Yang<sup>2</sup>, Shuang Yang<sup>2</sup>, Lars Bolund<sup>2,16</sup>, Karsten Kristiansen<sup>17</sup>, Hancheng Zheng<sup>2,18</sup>, Shaochuan Li<sup>2,18</sup>, Xiuqing Zhang<sup>2</sup>, Huanming Yang<sup>2</sup>, Jian Wang<sup>2</sup>, Rifei Sun<sup>1</sup>, Baoxi Zhang<sup>1</sup>, Shuzhi Jiang<sup>1</sup>, Jun Wang<sup>2,17</sup>, Yongchen Du<sup>1</sup> & Songgang Li<sup>2</sup>

**Cucumber is an economically important crop as well as a model system for sex determination studies and plant vascular biology. Here we report the draft genome sequence of *Cucumis sativus* var. *sativus* L., assembled using a novel combination of traditional Sanger and next-generation Illumina GA sequencing technologies to obtain 72.2-fold genome coverage. The absence of recent whole-genome duplication, along with the presence of few tandem duplications, explains the small number of genes in the cucumber. Our study establishes that five of the cucumber's seven chromosomes arose from fusions of ten ancestral chromosomes after divergence from *Cucumis melo*. The sequenced cucumber genome affords insight into traits such as its sex expression, disease resistance, biosynthesis of cucurbitacin and 'fresh green' odor. We also identify 686 gene clusters related to phloem function. The cucumber genome provides a valuable resource for developing elite cultivars and for studying the evolution and function of the plant vascular system.**

The botanical family Cucurbitaceae, commonly known as cucurbits and gourds, includes several economically important cultivated plants, such as cucumber (*C. sativus* L.), melon (*C. melo* L.), watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai) and squash and pumpkin (*Cucurbita* spp.). Agricultural production of cucurbits uses 9 million hectares of land and yields 184 million tons of vegetables, fruits and seeds annually (<http://faostat.fao.org>). The cucurbit family also displays a rich diversity of sex expression, and the cucumber has served as a primary model system for sex determination studies<sup>1</sup>. The cucurbits are also model plants for the study of vascular biology, as both xylem and phloem sap can be readily collected for studies of long-distance signaling events<sup>2,3</sup>.

Despite the agricultural and biological importance of cucurbits, knowledge of their genetics and genome is currently very limited. We have therefore sequenced and assembled the genome of the domestic cucumber, *C. sativus* var. *sativus* L.

All previous plant genome sequences have been derived using traditional Sanger technology<sup>4-9</sup>. The recent development of

<sup>1</sup>Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. <sup>2</sup>BGI-Shenzhen, Shenzhen, China. <sup>3</sup>Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark. <sup>4</sup>Department of Plant Biology, College of Biological Sciences, University of California, Davis, California, USA. <sup>5</sup>College of Life Sciences, Beijing Normal University, Beijing, China. <sup>6</sup>National Maize Improvement Center of China, Key Laboratory of Crop Genetic Improvement and Genome of Ministry of Agriculture, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China. <sup>7</sup>Boyce Thompson Institute and USDA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA. <sup>8</sup>High-Tech Research Center, Shandong Academy of Agricultural Sciences, Jinan, China. <sup>9</sup>US Department of Agriculture, Agricultural Research Service, Vegetable Crops Research Unit, Department of Horticulture, University of Wisconsin, Madison, Wisconsin, USA. <sup>10</sup>Diversity Arrays Technology, Canberra, Australia. <sup>11</sup>Wageningen UR Plant Breeding, Wageningen, The Netherlands. <sup>12</sup>The Graduate University of Chinese Academy of Sciences, Beijing, China. <sup>13</sup>High School Affiliated to Renmin University of China, Beijing, China. <sup>14</sup>Division of Applied Life Science (BK21 and WCU program), PMBBRC and EB-NCRC, Gyeongsang National University, Jinju, Republic of Korea. <sup>15</sup>National Engineering Research Center for Vegetables, Beijing, China. <sup>16</sup>Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. <sup>17</sup>Department of Biology, University of Copenhagen, Copenhagen, Denmark. <sup>18</sup>South China University of Technology, Guangzhou, China. <sup>19</sup>These authors contributed equally to this work. Correspondence should be addressed to Y.D. ([yongchen.du@mail.caas.net.cn](mailto:yongchen.du@mail.caas.net.cn)), S.H. ([huangsanwen@caas.net.cn](mailto:huangsanwen@caas.net.cn)), Jun Wang ([wangji@genomics.org.cn](mailto:wangji@genomics.org.cn)) or Songgang Li ([lisg@genomics.org.cn](mailto:lisg@genomics.org.cn)).

next-generation sequencing technologies has significantly improved sequencing throughput at a markedly reduced cost<sup>10</sup>. However, an intrinsic characteristic of next-generation technologies is their short read length (~50 bp), which prevents their direct application for *de novo* assembly of large genomes. When using these new technologies, assembly is typically carried out by mapping these short reads onto a known reference genome<sup>11,12</sup>. For the cucumber genome, we carried out a novel combination *de novo* sequencing strategy, taking advantage of the long read and clone length of Sanger technology and, for the first time, the high sequencing depth and low unit cost of Illumina GA technology.

## RESULTS

### Sequencing and assembly

We selected the 'Chinese long' inbred line 9930, which is commonly used in modern cucumber breeding<sup>13</sup>, for our genome sequencing project. We generated a total of 26.5 billion high-quality base pairs, or 72.2-fold genome coverage, of which the Sanger reads provided 3.9-fold coverage and the Illumina GA reads provided 68.3-fold coverage (Supplementary Table 1). The GA reads ranged in length from 42 to 53 bp.

We compared the assemblies obtained by Sanger reads only, Illumina GA reads only and Sanger plus Illumina reads. The 'hybrid' approach achieved markedly longer N50 (the size above which half of the total length of the sequence set can be found) in both contigs and scaffolds, so we used this assembly for further analyses (Table 1 and Supplementary Table 2). The total length of the assembled genome was 243.5 Mb, about 30% smaller than the genome size estimated by flow cytometry of isolated nuclei stained with propidium iodide (367 Mb)<sup>14</sup> and by *K*-mer depth distribution of sequenced reads (350 Mb; Supplementary Fig. 1). Several types of satellite sequences were present in the data set, comprising 23.2% of all Sanger reads and 76.2% of unassembled reads (Supplementary Table 3). FISH analysis indicated that these are primarily located in the centromeric and telomeric regions<sup>15</sup>. The cucumber genome also contains a large number of rRNA sequences, and about 3.3% of the Sanger reads matched 45S rRNA. These results indicated that the majority of the remaining 30% of unassembled regions of the genome are likely to be heterochromatic satellite or rRNA sequences.

The high coverage of the cucumber genome by this assembly was also confirmed using the available EST, fosmid and BAC sequences. The assembly contains 96.8% of the 63,312 cucumber unigenes assembled from ~350,000 Roche 454-sequenced ESTs, 99.3% of the 6,952 NCBI-deposited ESTs of cucumber, 91.2% of the 50,441 NCBI-deposited ESTs of melon and 98.7% of the six finished fosmid and BAC sequences (Supplementary Table 4).

A genetic map was developed using 77 recombinant inbred lines from the intersubspecific cross between Gy14 (a North American processing market-type cucumber cultivar) and PI183967 (an accession of *C. sativus* var. *hardwickii* originating from India). The map spans 581 cM and contains 1,885 markers, including 995 micro-satellite markers<sup>16</sup> and 890 Diversity Arrays Technology markers (marker sequences can be accessed at <http://cucumber.genomics.org.cn>). Using this map, we were able to anchor 72.8% of the assembled sequences onto the seven chromosomes. Among the 1,885 markers, 1,763 (93.5%) were uniquely aligned and used for constructing the pseudochromosomes. The majority (98.7%) of the markers were collinear with the sequence assembly (Fig. 1a). Comparison of the genetic and physical distances between markers revealed

**Table 1 Cucumber genome assembly statistics**

Assembly	Contig N50 <sup>a</sup> (kb)	Contig total (Mb)	Scaffold N50 (kb)	Scaffold total (Mb)	% sequence anchored on chromosome
Sanger	2.6	204	19	238	—
Illumina GA	12.5	190	172	200	—
Sanger + Illumina GA	19.8	226.5	1,140	243.5	72.8%

<sup>a</sup>N50 refers to the size above which half of the total length of the sequence set can be found.

recombination suppression of two 10-Mb regions at either end of chromosome 4, a 20-Mb region on chromosome 5 and an 8-Mb region on chromosome 7. Using high-resolution FISH, we confirmed previously identified segmental inversion<sup>16</sup> within the suppression region on chromosome 5 between Gy14 and PI183967 (Fig. 1b), which provides an explanation for recombination suppression in these regions. These regions of recombination suppression are additionally useful for studying cucumber evolution during domestication.

After excluding 16 markers whose genetic positions were ambiguous, we examined the six remaining regions that had conflicts between the genetic map and our assembly. Upon inspection, we found that clone mate-pair information supported our assembly in all of these regions (Supplementary Fig. 2). We also identified no misassembly within the regions covered by the six finished fosmid or BAC sequences (Supplementary Fig. 3). The conflicts may be a result of chromosomal rearrangement that occurred between the sequenced genotype 9930 and the genotypes used to create the mapping population; alternatively, these markers may have been placed incorrectly on the genetic map. Sequencing depth distribution showed that we obtained more than 10× coverage on more than 97.5% of the assembly (Supplementary Fig. 4).

### Repetitive sequences and transposons

The cucumber genome contains a large number of transposable elements, but only a few have previously been identified. We therefore constructed repeat libraries using multiple *de novo* methods and then derived a combined repeat library that contained 1,566 sequences (Supplementary Table 5), of which 469 (29.9%) were manually classified (Supplementary Table 6). We then used this library for repeat annotation of the cucumber genome. We identified a total of 54.4 Mb, which represents ~24% of the genome, as repeats. Among them, 51.5% could be classified based on known repeats. The long terminal repeat (LTR) retrotransposons (*gypsy* and *copla*) made up the majority of the transposable element classes and comprised 10.4% of the genome (Supplementary Table 7). The repeats divergence rate (percentage of substitutions in the matching region compared with consensus repeats in constructed libraries) distribution showed a peak at 20%. A fraction of LTR retrotransposons, long interspersed nuclear elements and DNA transposons (composing 2.3%, 0.4% and 0.2% of the genome, respectively) are of relatively recent origin, having a sequence divergence rate of less than 5% (Supplementary Fig. 5).

### Gene annotation

We used three gene-prediction methods (cDNA-EST, homology based and *ab initio*) to identify protein-coding genes and then built a consensus gene set by merging all of the results (Supplementary Fig. 6). We predicted 26,682 genes, with a mean coding sequence size of 1,046 bp and an average of 4.39 exons per gene (Supplementary Table 8). Under an 80% sequence overlap threshold, we found that 26.7% of the genes were supported by models from all three gene prediction methods, 25% had both *ab initio* prediction and homology-based evidence, and 7.4% had *ab initio* prediction and cDNA-EST expression evidence; the remaining genes were primarily derived from pure

**Figure 1** Integrated genetic and physical map of cucumber. **(a)** Genetic versus physical distance map of the seven cucumber chromosomes. The genetic map was constructed using a recombinant inbred line mapping population from the intersubspecific cross between Gy14 (domestic cucumber) and PI183967 (wild cucumber). **(b)** Segmental inversion between Gy14 and PI183967 on cucumber chromosome 5 detected by high-resolution FISH (12-2 and 12-7 denote individual fosmid clones). A low-resolution FISH analysis was also recently reported<sup>16</sup>. Scale bars represent 1  $\mu$ m.

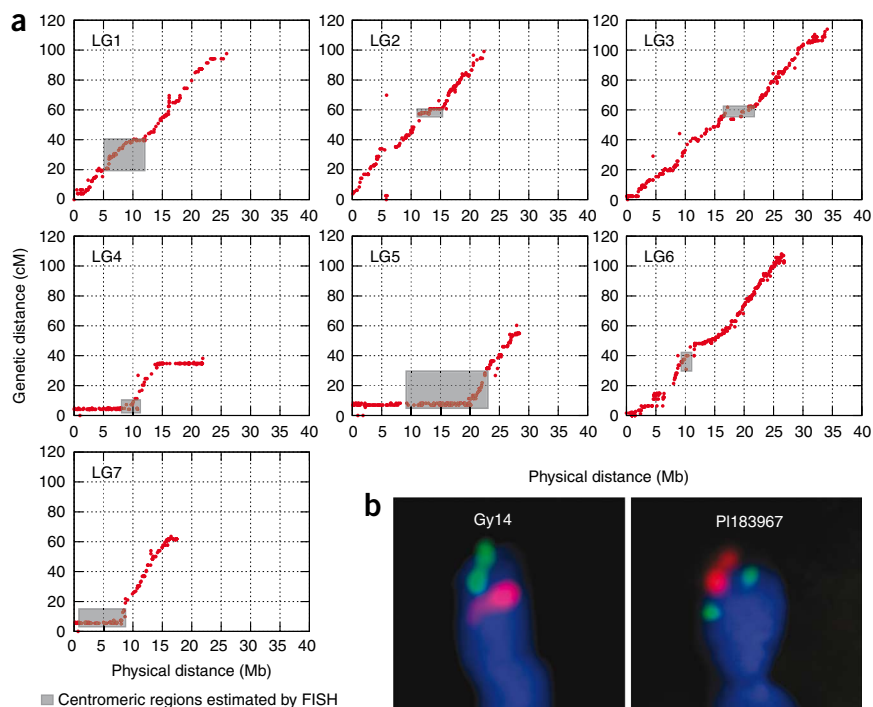
*ab initio* prediction, but the majority of these were supported by multiple gene finders (Supplementary Table 9). About 81% of the genes have homologs in the TrEMBL protein database, and 66% can be classified by InterPro. In sum, 82% of the genes have either known homologs or can be functionally classified (Supplementary Table 10). In addition to protein-coding genes, we identified 292 rRNA fragments and 699 tRNA, 238 small nucleolar RNA, 192 small nuclear RNA and 171 miRNA genes in the cucumber genome (Supplementary Table 11).

On the basis of pairwise protein sequence similarities, we carried out a gene family clustering analysis on all genes in sequenced plants, using rice as an outgroup. The cucumber genes consist of 15,669 families. Of these, 4,362 are cucumber unique families, among which 3,784 are single-gene families (Supplementary Table 12). The EST confirmation rate of these unique single-copy genes was much lower than the average of all predicted genes (33.4% vs. 72.3%, respectively). This category may therefore contain a number of false-positive predictions. In papaya, there are 4,622 unique families, but the actual number of genes is estimated to be 24,746, which is lower than the 28,629 predicted genes<sup>7</sup>. Thus, the actual number in cucumber should be lower than 26,682 and similar to that in papaya. The smaller average gene family size in cucumber (1.71) and papaya (1.77) supports this conclusion (Fig. 2a).

The cucumber genome contains the smallest number of tandem gene duplications (479) among all the plants we compared, whereas grapevine has the largest number (5,382; Fig. 2a). This may contribute in part to the small number of genes in cucumber.

### Absence of recent whole-genome duplication

Whole-genome duplication (WGD) is common in angiosperm plants and produces a tremendous source of raw material for gene genesis. Previous research has revealed a paleohexaploidy ( $\gamma$ ) event in the common ancestor of *Arabidopsis thaliana* and grapevine after the divergence of monocotyledons and dicotyledons<sup>6</sup>. Subsequently, two WGDs ( $\alpha$  and  $\beta$ ) occurred in *Arabidopsis*<sup>17</sup> and one ( $p$ ) in poplar<sup>8</sup>, whereas no recent WGD occurred in grapevine and papaya. Evidence indicates that rice underwent an ancient WGD<sup>18</sup>. We carried out a collinear gene-order analysis on the cucumber genome and observed no recent WGD and only a few segmental duplication events (Supplementary Fig. 7). We also used the distance-transversion rate at fourfold degenerate sites (4DTV method) to analyze paralogous gene pairs between syntenic blocks in *Arabidopsis* and cucumber, respectively. Two peaks ( $\sim 0.06$  and  $\sim 0.25$ ) in *Arabidopsis* support the



two recent WGDs (Fig. 2b). In cucumber, the analysis showed ancient duplication events (peak at  $\sim 0.60$ ) but did not reveal recent WGD. This lack of recurrent WGD in the small cucumber genome provides an important complement to the grapevine and papaya genomes to study ancestral forms and arrangements of plant genes.

### Syntenic with flowering plant genomes

Given the similar gene arrangements between cucumber and other plant genomes, we defined syntenic blocks that contained 5,473, 6,525, 9,842, 8,439 and 3,992 cucumber genes collinear to *Arabidopsis*, papaya, poplar, grapevine and rice, respectively (Supplementary Table 13 and Supplementary Figs. 8–12). The numbers of collinear genes were consistent with the phylogenetic distances of the other plants to cucumber. Within the syntenic blocks, we observed the highest density of collinear genes between cucumber and grapevine (90.5 genes per Mb), followed by papaya (76.1; the low contiguity of genome assembly may have, in part, decreased this value), poplar (68.8), rice (55.6) and *Arabidopsis* (43.5; Supplementary Table 13). This indicates that *Arabidopsis* has the most reshuffled or rearranged genome, whereas the genomes of grapevine and papaya are more conserved, probably because they have not undergone WGD since the ancestral paleohexaploidy.

### Substantial fusion events involved in chromosomal evolution

Melon and cucumber belong to the same genus, although cucumber has seven chromosomes and melon has 12. Watermelon, their common distant relative, has 11 chromosomes. To investigate cucurbit chromosomal evolution, we compared the melon<sup>19</sup> and watermelon genetic maps to the cucumber genome (Fig. 3a). In total, 348 (66.7%) of the 522 melon markers and 136 (58.6%) of the 232 watermelon markers were aligned on the cucumber chromosomes











